Research Methodology - Assignment 2

Jos Elkink
jos.elkink@ucd.ie

http://jaeweb.cantr.net/teaching/research-methodology

Fall 2007

Download the War Vote dataset posted under lecture 3 and open it in R:

```
library(foreign)
d <- read.csv(file.choose())
names(d)
summary(d)
attach(d)
```

The data consists of the following variables: (1) the name of the state, (2) the number of soldiers from that state killed in the Iraq War up to that point, (3) the state population, (4) the state's median family income, (5) the percentage of people in the state with a college education, (6) the unemployment rate in the state, (7) the per capita amount of federal aid distributed to the state, (8) the percentage of the state population that is white, (9) the percentage of people in the state that own their own homes, (10) the per capita energy consumption rate (in British Thermal Units), and (11) the percentage of voters in the state who voted for George W. Bush. These 11 variables are observed in all 50 US states and the District of Columbia just before the 2004 Presidential election.

We are interested in the following hypothesis: "States that receive more money from the federal government will support the incumbent president at higher rates."

Write an essay of approximately 700 words plus graphs or tables as necessary, in which you answer the following questions:

1) What variable in the data should you use for your dependent variable? What variable should you use for your independent variable?

2) Using a histogram, examine your dependent and independent variables. In addition, use the following code[1] to generate a scatter plot with ID numbers for each case:

[1] Note that no explicit d$ is used to use the variable names because the attach() command was used.

```
plot(bush.vote ~ fed.aid, data=d, cex=0)
text(fed.aid, bush.vote, 1:length(fed.aid))
```

Which observation or observations are outliers in your dependent variable? Which observation or observations are outliers in your independent variable?

If a case is a bad outlier, you might want to remove it from your analysis. Do not make an automatic habit of this, though - for more advanced analysis, it is much better to understand what variable you are missing that explains why this case is an outlier. If you decide, however, to remove, for example, the 20th observation, you can do the following:

```
detach(d)
d2 <- d[-20,]
attach(d2)
```

3) Calculate the correlation between your dependent and independent variables. What is this correlation? Based on this correlation, does it appear there is support for the hypothesis?

4) Run a regression to examine the relationship between the dependent and independent variables. What is the intercept of the regression line, and what does it tell you in this case? What is the slope of the regression line, and what does it tell you in this case? Based on this regression line, does there appear to be support for the hypothesis?

5) Generate a scatterplot with a regression line showing the relationship between your dependent and independent variables (make sure your dependent variable is on the vertical axis).

6) Make up a hypothesis of your own that you can study with this data. What is the hypothesis? And what dependent and independent variables does it involve?

7) Study this hypothesis in a similar vain as the one above. What is your conclusion?

To run a regression and add the regression line to a scatterplot, you can use the following code (replace x and y with the independent and dependent variables, respectively, of interest):

```
lm(y ~ x)
plot(y ~ x, ylab="Y variable", xlab="X variable", pch=19)
abline(lm(y ~ x), lwd=2)
```

Whereby the ylab and xlab options allow you to label the axes, the lwd option alters the thickness of the line, and the pch option alters the look of the dots.

This assignment is due by email on Friday, October 12, 5 pm.